Lecture 7: Data Center Networks

CSE 222A: Computer Communication Networks

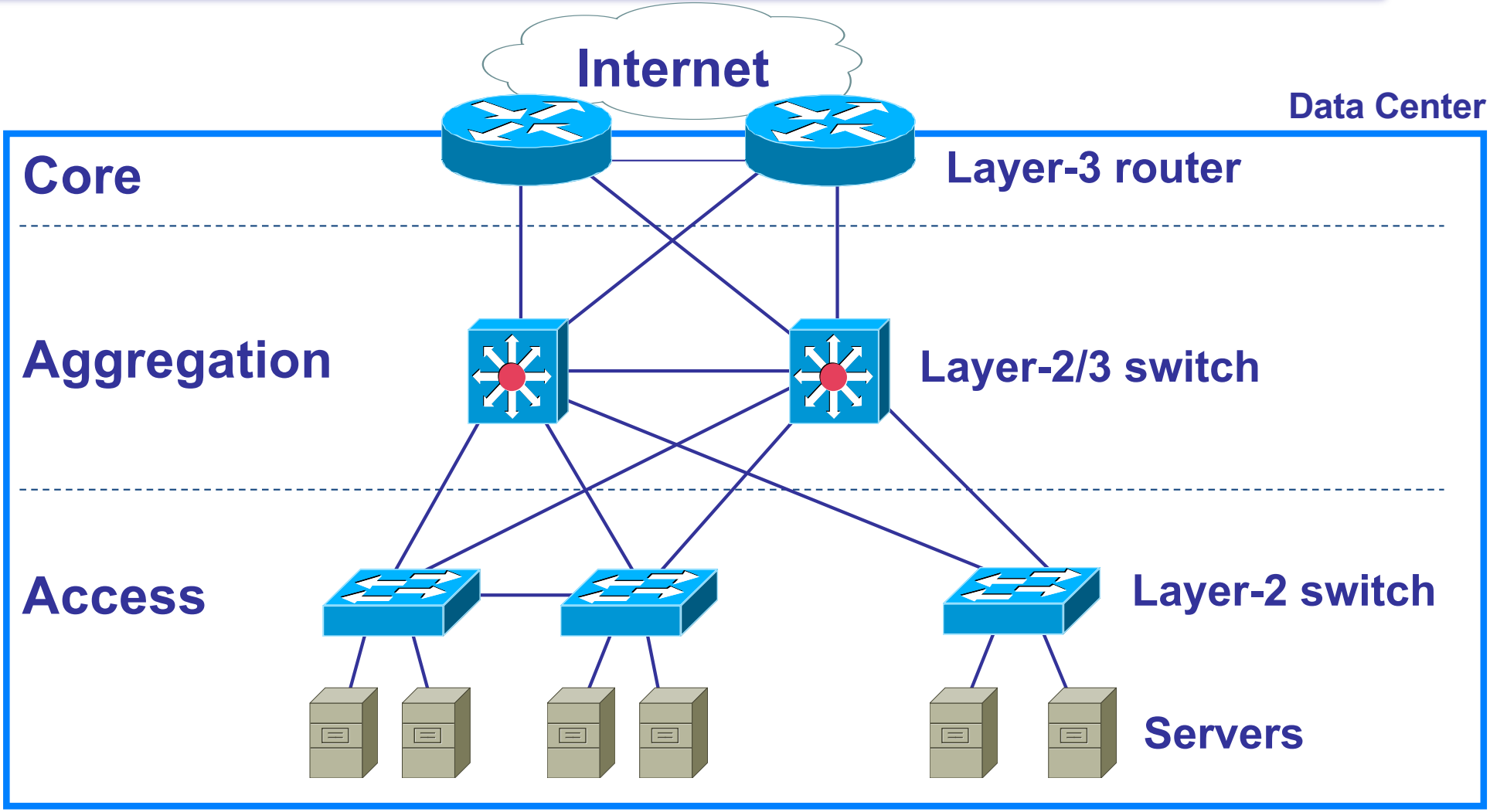
Alex C. Snoeren

Thanks: Nick Feamster

Lecture 7 Overview

- Project discussion
- Data Centers overview
- “Fat Tree” paper discussion

Traditional DC Topology



DC Network Requirements

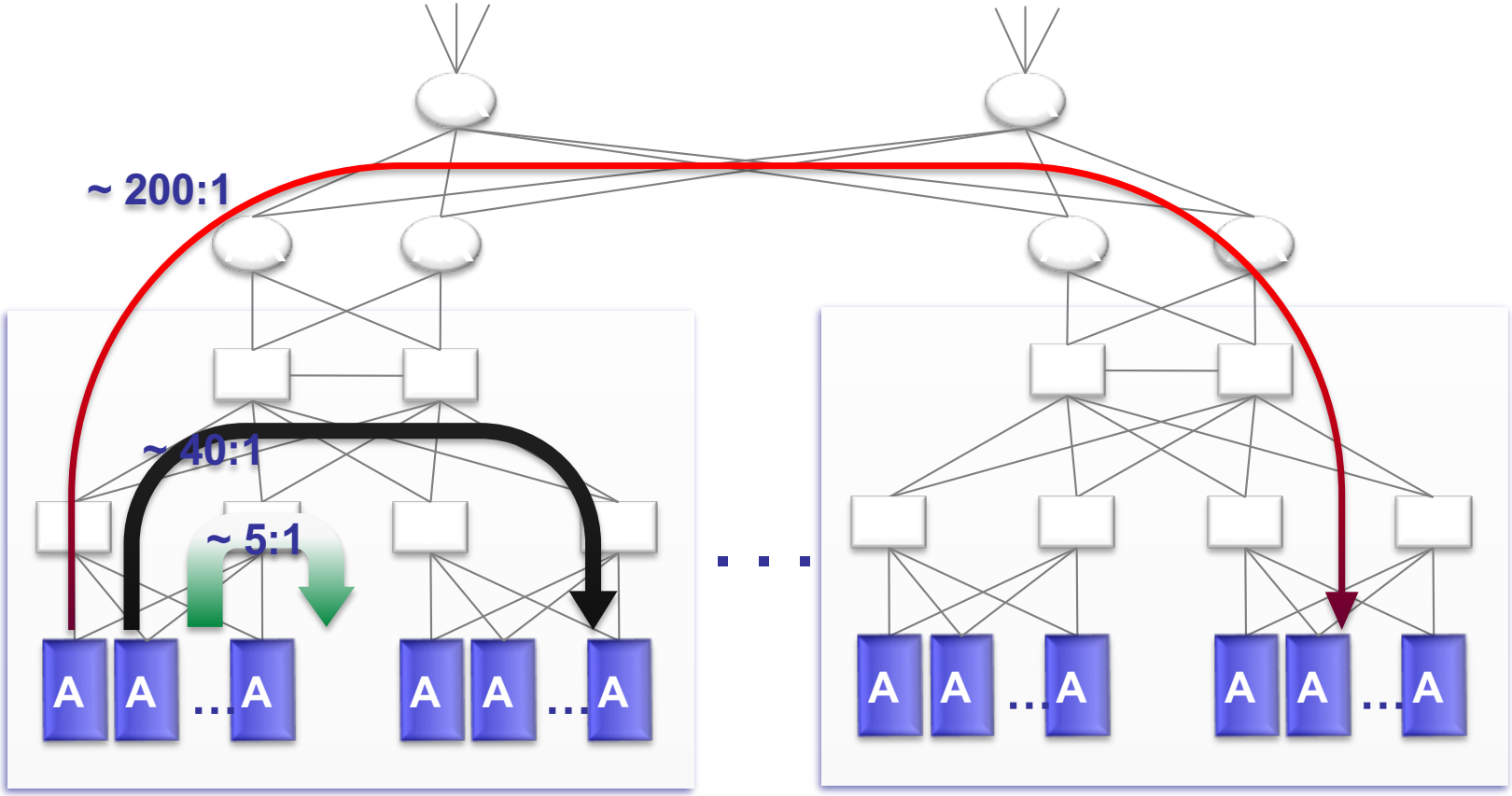
- Scalability
  - Incremental build out?
- Reliability
  - Loop free forwarding
- VM migration

- Reasonable management burden
  - Humans in the loop?

Traditional Topologies

- Over subscription of links higher up in the topology
- Tradeoff between cost and provisioning
- Single point of failure

Capacity Bottlenecks



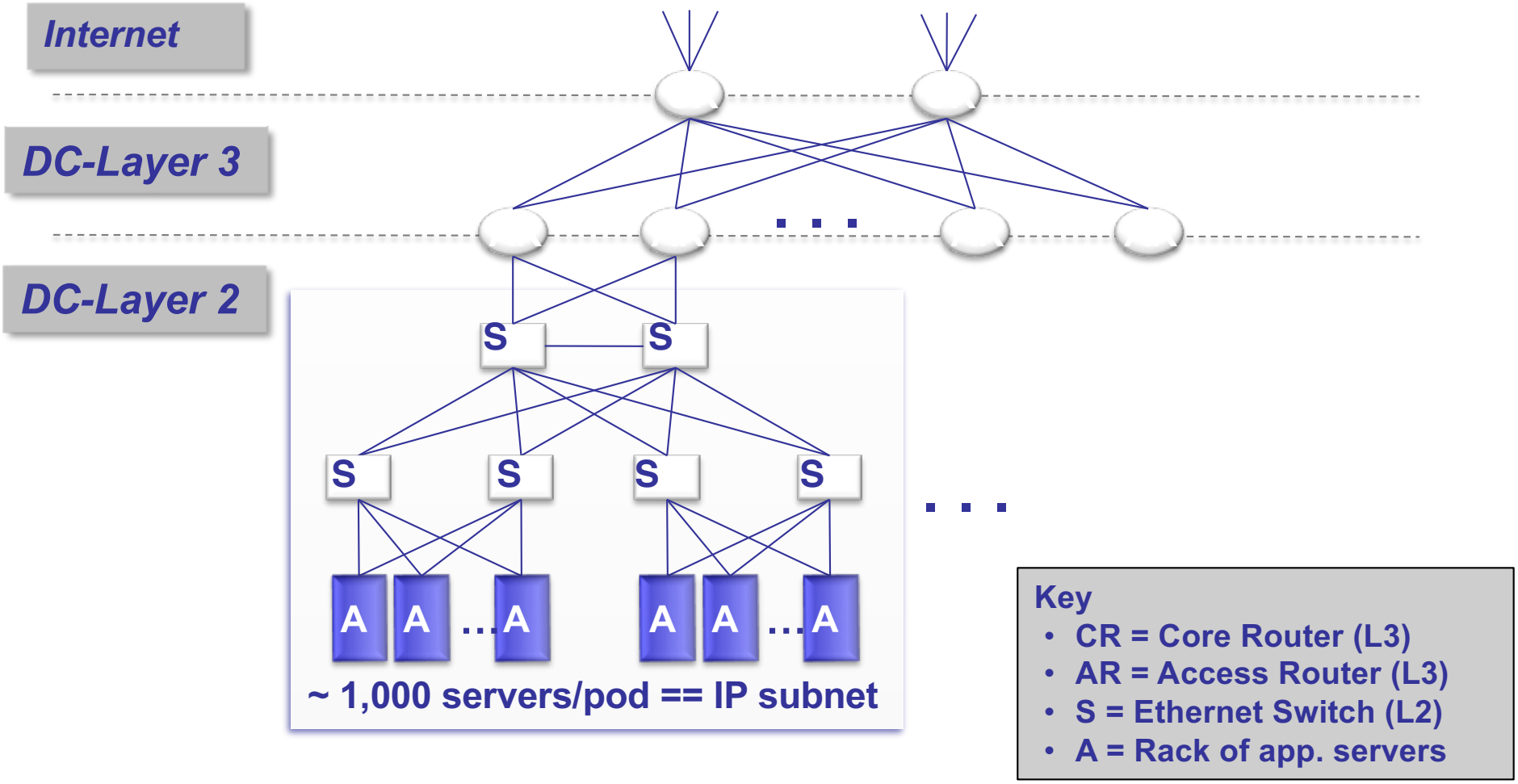
Management: L2 vs. L3

- Ethernet switching (layer 2)
  - Cheaper switch equipment
  - Fixed addresses and auto-configuration
  - Seamless mobility, migration, and failover
- IP routing (layer 3)
  - Scalability through hierarchical addressing
  - Efficiency through shortest-path routing
  - Multipath routing through equal-cost multipath
- Data centers often connect layer-2 islands by IP routers

Advantages of Layer 2

- Certain monitoring apps require server with same role to be on the same VLAN
- Using same IP on dual homed servers
- Allows organic growth of server farms
- VM migration is easier

Layer 2 Pods w/L3 Backbone

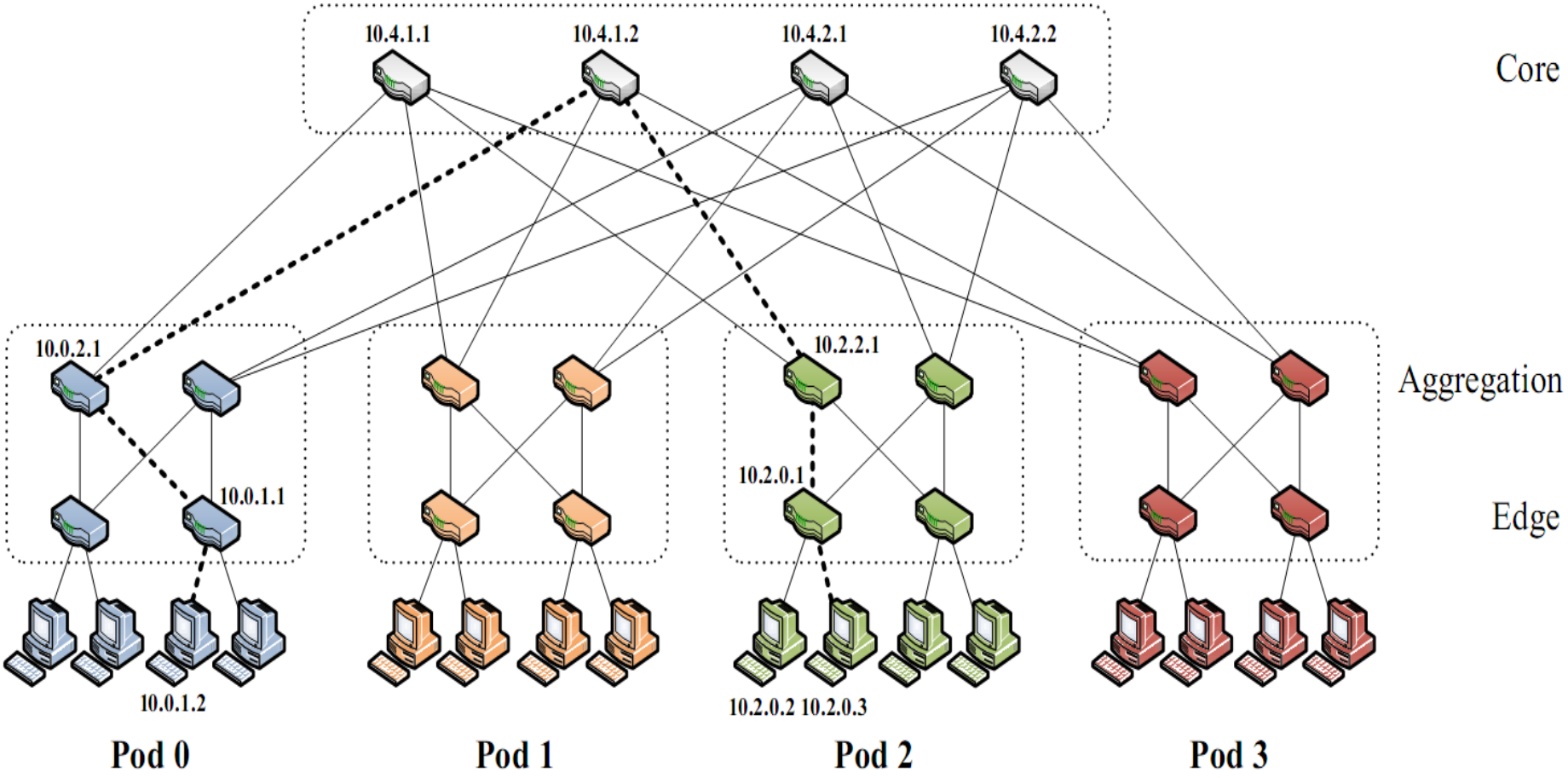


FAT Tree-Based Solution

- An all Layer-3 solution

- Connect end-host together using a “fat-tree” topology
  - Infrastructure consist of cheap devices
    - » Each port supports same speed as endhost
  - All devices can transmit at line speed if packets are distributed along existing paths
  - A k-port fat tree can support $k^3/4$ hosts

“Fat-Tree” Topology



Fat-Tree Challenges

- Layer 3 will only use one of the existing equal cost paths
- Packet re-ordering occurs if layer 3 blindly takes advantage of path diversity
  - E.g., ECMP

Modified Fat Tree

- Enforce special addressing scheme in DC
  - Allows host attached to same switch to route only through switch
  - Allows inter-pod traffic to stay within pod
  - unused.PodNumber.switchnumber.Endhost
- Use two level look-ups to distribute traffic and maintain packet ordering.

Two-Level Lookups

- First level is prefix lookup
  - Used to route down the topology to endhost
- Second level is a suffix lookup
  - Used to route up towards core
  - Diffuses and spreads out traffic
  - Maintains packet ordering by using the same ports for the same endhost

Prefix	Output port
10.2.0.0/24	0
10.2.1.0/24	1
0.0.0.0/0	

→

Suffix	Output port
0.0.0.2/8	2
0.0.0.3/8	3

Diffusion Optimizations

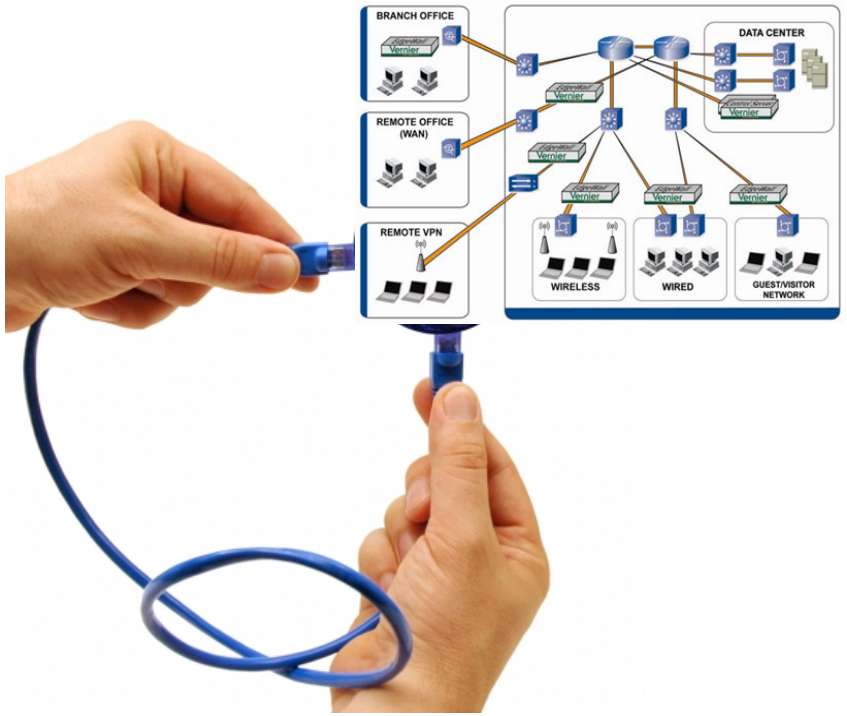
- Flow classification
  - Eliminates local congestion
  - Assign to traffic to ports on a per-flow basis instead of a per-host basis
- Flow scheduling
  - Eliminates global congestion
  - Prevent long lived flows from sharing the same links
  - Assign long lived flows to different links

Discussion

- VM mobility
- Data center build out
- Internet connectivity
- Address space allocation
  - Doing flow placement anyway?

The Beauty of Ethernet

- Flat addressing and self-learning enables plug-and-play networking
- Permanent and location independent addresses also simplify
  - Host mobility
  - Access-control policies
  - Network troubleshooting



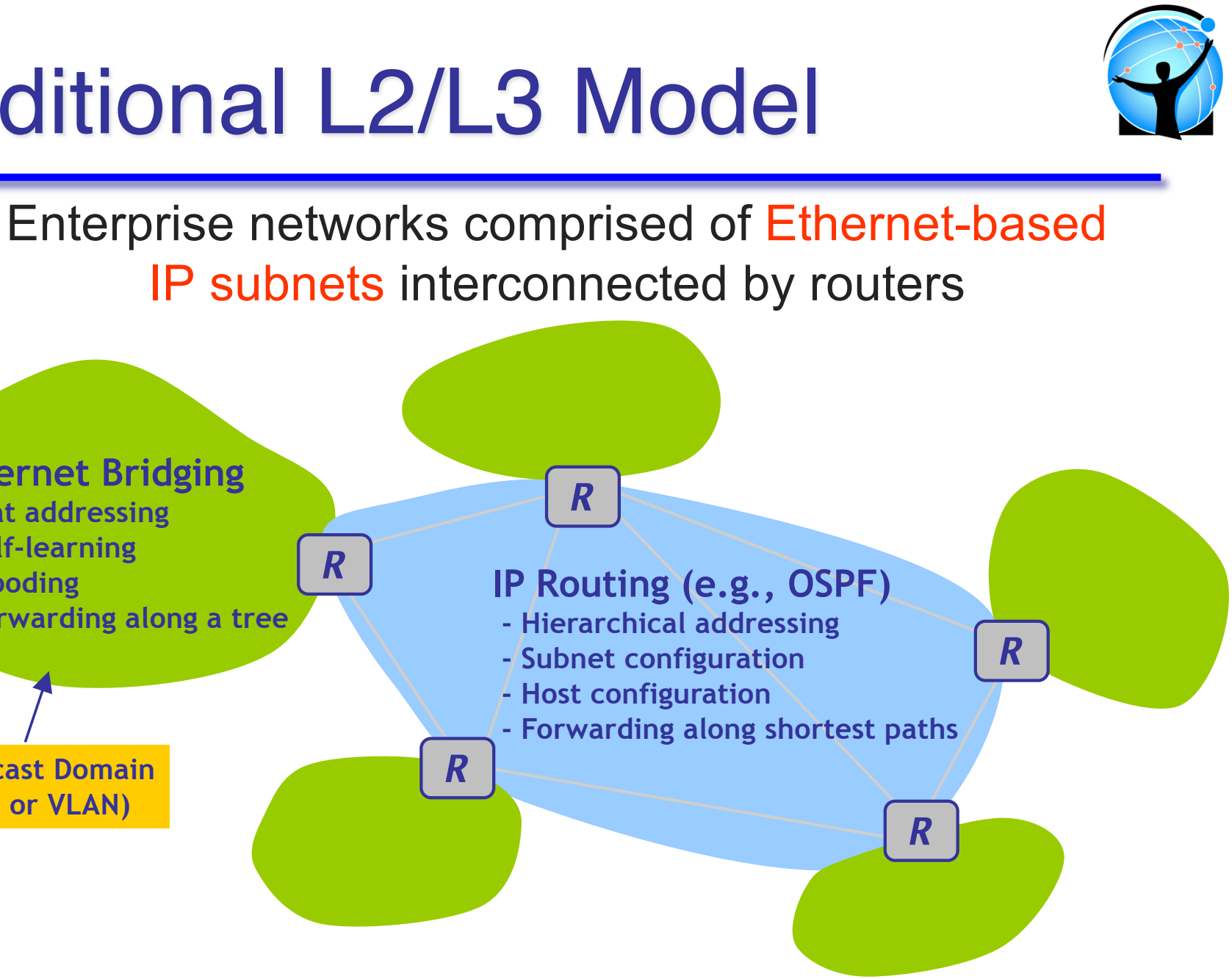
Ethernet Scaling Challenges

- Flooding-based delivery
  - Frames to unknown destinations are flooded

- Broadcasting for basic service
  - Bootstrapping relies on broadcasting
  - Vulnerable to resource exhaustion attacks

- Inefficient forwarding paths
  - Loops are fatal due to broadcast storms; uses STP
  - Forwarding along a single tree leads to inefficiency and lower utilization

Traditional L2/L3 Model



Enterprise networks comprised of Ethernet-based IP subnets interconnected by routers

For Next Class…

- Read and review PortLand paper
- Work on project proposals
  - Details available on the course Web page